A Unified Probabilistic Formulation of Image Aesthetic Assessment

Hui Zeng, Zisheng Cao, Lei Zhang, *Fellow, IEEE* and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—Image aesthetic assessment (IAA) has been attracting considerable attention in recent years due to the explosive growth of digital photography in Internet and social networks. The IAA problem is inherently challenging, owning to the ineffable nature of the human sense of aesthetics and beauty, and its close relationship to understanding pictorial content. Three different approaches to framing and solving the problem have been posed: binary classification, average score regression and score distribution prediction. Solutions that have been proposed have utilized different types of aesthetic labels and loss functions to train deep IAA models. However, these studies ignore the fact that the three different IAA tasks are inherently related. Here, we reveal that the use of the different types of aesthetic labels can be developed within the same statistical framework, which we use to create a unified probabilistic formulation of all the three IAA tasks. This unified formulation motivates the use of an efficient and effective loss function for training deep IAA models to conduct different tasks. We also discuss the problem of learning from a noisy raw score distribution which hinders network performance. We then show that by fitting the raw score distribution to a more stable and discriminative score distribution, we are able to train a single model which is able to obtain highly competitive performance on all three IAA tasks. Extensive qualitative analysis and experimental results on image aesthetic benchmarks validate the superior performance afforded by the proposed formulation. The source code is available at https://github.com/HuiZeng/Unified_IAA.**

***Index Terms*—Image aesthetic assessment, unified probabilistic formulation.**

I. Introduction

Given the explosive growth of digital photography in Internet and social networks, automatic image aesthetic assessment (IAA) has become increasingly important in many applications such as image search, personal photo album curation, photographic composition and image aesthetic manipulation [8, 19, 25, 30, 47]. Although it has been shown that there exist certain connections between human aesthetic experience and long-established photographic rules of thumb [2], it remains a very challenging task to computationally model image aesthetics, because of the high degree of implied subjectivity and the complex relationship that exists among many intertwined factors, contributing to the aesthetic visual sense [9, 19].

H. Zeng and L. Zhang are with the Department of Computing, The Hong Kong Polytechnic University. (E-mail: cshzeng@comp.polyu.edu.hk, cslzhang@comp.polyu.edu.hk).

Zisheng Cao is from the Da-Jiang Innovations (DJI). (E-mail: zisheng.cao@dji.com).

Alan C. Bovik is with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (E-mail: bovik@ece.utexas.edu).

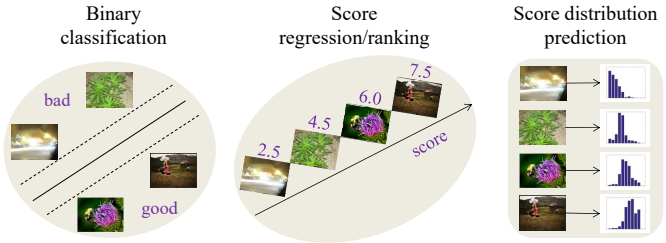


Fig. 1. Illustrations of the three most common tasks that image aesthetic assessment engines are trained to conduct.

Three principal approaches to the IAA problem have been undertaken: binary classification, quality score regression and score distribution prediction. Illustrations of these three tasks are shown in Fig. 1. While it is an intuitive idea to assign a single aesthetic score to a given image using a trained model, earlier attempts have revealed that accurate aesthetic score regression is a very challenging task. To make the problem more tractable, Datta *et al* [7] and Ke *et al* [22] formulated IAA as a binary classification problem where the goal is only to discriminate between aesthetically pleasing (or professional) photos from displeasing (amateur) photos, while ignoring images of ordinary aesthetic quality. Since then, a number of studies have been reported that follow a similar binary classification pipeline [3, 28–32, 34, 45].

However, as pointed out in [24], defining such a binary classification task in this context is questionable since there exists no clear boundaries between the aesthetic quality levels of images. For example, in the AVA dataset [37] which currently is the largest IAA subjective benchmark, the subjective scores assigned to all of the images in the corpus are approximately Gaussian distributed. This lack of multiple aesthetic modes implies that it is inappropriate to divide the images into two classes using a hard threshold score. Predicting numerical aesthetic quality scores or relative aesthetic rankings can likely provide rich and more useful information than binary classification over a broader range of practical IAA applications [10, 39], although accurately regressing on aesthetic quality scores is a much more complex task.

Predicting an aesthetic score distribution has recently begun to attract attention as an even more descriptive alternative approach [17, 18, 36, 45, 46]. A common argument that is used to motivate these studies is that a single scalar score is not sufficiently informative to describe the subjective nature of image aesthetics. These authors instead suggest predicting the distributions of scores that are collected from human raters, who are assumed to supply the best possible information

regarding image aesthetics. However, raw score distributions are not entirely reliable and contain noise [18, 45, 46], and moreover there is no widely accepted metric for evaluating the performances of score distribution predictors. Because of this, most of these studies operate by ultimately converting the predicted score distribution to either a scalar aesthetic quality score or binary aesthetic label.

As can be seen, the three tasks utilize different kinds of aesthetic labels. Previous studies have treated them as different problems, by designing various architectures, and by employing different loss functions and different evaluation metrics. However, these studies ignore the fact that the three IAA tasks are inherently related, which has been rarely discussed in the literature. Here, we make the first attempt, to the best of our knowledge, to unfold the relationships that exist between the three different IAA tasks, and we provide a unified formulation to handle all of them. Specifically, we show that the three types of aesthetic labels (binary category labels, quality scores, and score distributions) can be transformed into a unified probabilistic formulation. This unified formulation further motivates us to employ a more efficient and effective loss function to train deep IAA models. More importantly, we are able to use a single model trained on the score distribution to solve all three tasks. Such an approach is more efficient and lightweight in practice than using three different models to conduct different IAA tasks. Since the raw score distributions are quite noisy, limiting the performance of models trained on them, we propose a way to process the raw data to obtain a more stable and discriminative score distribution. Training deep IAA models on the stabilized score distribution produces more robust and accurate results. Finally, we are able to obtain highly competitive performance on all the three tasks using a single model.

Our main contributions are summarized as follows:

(1) We provide a unified probabilistic formulation that can be used to train a deep network to efficiently perform each of the three aforementioned IAA tasks: binary classification, score regression and score distribution prediction;
(2) By employing this unified probabilistic formulation, we are naturally able to introduce a more efficient and effective loss function when optimizing deep IAA models, which is applicable to all three tasks;
(3) We conduct an in-depth analysis of the noise contained in raw subjective score distributions and we propose a way of fitting the scores to a more stable and discriminative score distribution. A single model trained using the stabilized score distribution leads to state-of-the-art performances on all three IAA tasks.

II. Related work

Early work on the IAA problem [7, 22, 30, 34, 46] was mainly focused on the design of hand-crafted features, while recent ones [17, 18, 21, 24, 28, 29, 31, 32, 36, 45] generally employ deep models. In the following, we discuss the representative approaches to each of the IAA tasks, and refer interested readers to [9, 19] for more comprehensive reviews of the existing literature.

Binary classification. Most existing IAA models have been designed for the binary classification task. This is mainly because early pioneering attempts [7, 22] found it difficult to accurately predict aesthetic quality scores using low-level, hand-crafted, and often heuristic features. Early on, Datta *et al.* [7] modeled image aesthetics by computing a set of low-level visual features, and using them to classify aesthetically pleasing and displeasing images, and to regress on numerical aesthetic scores. By analyzing the observable differences between professional and amateur photos, Ye *et al.* [22] designed a set of high-level semantic features, achieving promising binary classification accuracy on the CUHK dataset [22]. Luo *et al.* [30] furthered progress in this direction by first segmenting potentially interesting subjects from the images then applying semantic features to significantly improve classification accuracy. Marchesotti *et al.* [34] employed very general local descriptors like SIFT [27] instead of carefully designed aesthetical features and encoded their distribution using a Bag-of-Visual-Words [5] and the Fisher Vector [40]. Their generic representation was also able to obtain competitive classification accuracy.

Recently, a number of deep models [28, 29, 31, 32] have been proposed for binary aesthetic classification. These efforts have mainly focused on extracting effective aesthetic features while optimizing the standard cross-entropy (CE) loss to train their deep models. Lu *et al.* [28] deployed independent parallel convolutional neural networks (CNNs) to model the overall layout and fine-grained details of image aesthetics. To alleviate biases caused by training on a single crop, Lu *et al.* [29] introduced a multi-patch aggregation network to better capture local aesthetic information. Mai *et al.* [32] employed an adaptive spatial pyramid pooling [13] strategy to preserve the aspect ratio and other compositional aspects of images, aggregating several sub-networks to leverage multi-scale information. Ma *et al.* [31] significantly improved performance using the multi-patch aggregation pipeline via an adaptive patch selection strategy. Sheng *et al.* [41] proposed a multi-patch aggregation model which adaptively adjusts the weight of each patch, based on an attention mechanism. Liu *et al.* [26] proposed a semi-supervised deep active learning algorithm combined with a computed gaze shifting path to achieve highly competitive classification accuracy on the AVA dataset.

Quality score regression. More recently, researchers have been able to obtain reasonable performance on the problem of regressing on aesthetic quality scores. Kong *et al.* [24] trained a deep model to regress on scalar aesthetic quality score and on the relative ranking order. They employed the mean square error (MSE) as the basic loss function, and used multi-task learning to further improve performance. A contemporary work [17] also employed the MSE loss to train a deep regression model. Similar work involving quality score regression has been done for many years in a closely related research area, image quality assessment (IQA) [6, 20, 23].

Score distribution prediction. Predicting aesthetic score distributions has attracted significant attention recently. It appears that Wu *et al.* [46] made the first attempt to employ a subjective score distribution to better represent image aesthetics. They proposed a structure regression algorithm to

model the score distribution, along with two learning strategies to handle unreliabilities in score distribution. Jin *et al.* [17] trained a deep CNN model to predict the aesthetic score distribution using the Chi-square distance and introduced a weighting strategy to address the sample imbalance problem. Jin *et al.* [18] employed the Jensen-Shannon divergence and trained a variety of deep IAA models by using several distance metrics to produce score distribution predictions. In an attempt to account for unreliability of the score distribution, Wang *et al.* [45] fitted a Gaussian distribution to the score distribution and trained a deep CNN model to simultaneously predict the low-order moments (mean and variance) of the distribution using the Kullback-Leibler (KL) divergence distance. Talebi *et al.* [43] employed the squared Earth Mover's Distance (EMD) [15] to train IAA models to leverage prior information regarding the ordering classes of aesthetic quality ratings.

Since the different types of aesthetic labels (binary labels, scalar quality scores and vectorized score distributions) have different formats and scales, previous approaches to the three different IAA subtasks have deployed a diversity of loss functions and evaluation metrics. Unlike all prior methods which have focused on a single specific task, using one specific loss function and one particular type of label to train a model, we advance progress on all three IAA tasks by revealing the close relationship that exists among them, and we also propose a unified formulation that is able to solve all three of them. Specifically, our "unified formulation" allows each of the three IAA tasks to be learned using the same loss function with corresponding labels. By contrast, the loss functions employed in previous models are too specific to be applied to all three types of aesthetic labels. For example, loss functions based on the cumulative distribution (such as EMD used in [15, 43]) or on high-order moments (such as variance, skewness and kurtosis) [18, 45] cannot be generalized to learn binary classification and quality score regression. In addition, we also present a unique analysis of noise in the raw score distribution, and we propose an effective stabilization strategy to ameliorate its effects.

Another interesting research problem is to predict the aesthetic perception of individual users. Recent studies [1, 48] of the subjective assessment of the aesthetic quality of paintings have shown that higher accuracies can be reached when treating observers individually. Unfortunately, all current publicly available image aesthetic datasets provide only the overall score distributions on each image collected through online crowdsourcing. The scores of individual observers are not available, preventing us from conducting such experiments.

III. Proposed method

A. Aesthetic Labels

Given a dataset with N sample images, assume that a set of rating scores $\mathbf{s}_n = \{s_n(1), s_n(2), ..., s_n(K)\}$ was collected from K different human raters on each image $\mathbf{x}_n$. Usually, the rating scales are predefined as a set of ordinal integers $\mathbf{b} = [b_1, b_2, ..., b_M]$, where each integer represents an aesthetic quality level while we will also assume that a larger value signifies a higher aesthetic quality. The ground truth score that associated with each image is usually defined as the average rating (or mean opinion) score:

$$\mu_{s_n} = \frac{1}{K}\sum_{k=1}^{K} s_n(k), \tag{1}$$

where μ_{s_n} is a continuous value in the range $[b_1, b_M]$.

By thresholding the average score, images are divided into two classes, where the class labels are assigned as:

$$c_n = \begin{cases} 1, \text{when } \mu_{s_n} > b_{\lceil \frac{M}{2} \rceil} + \delta, \\ 0, \text{when } \mu_{s_n} \leq b_{\lceil \frac{M}{2} \rceil} - \delta, \end{cases} \tag{2}$$

where $b_{\lceil \frac{M}{2} \rceil}$ is the mid-point of a pre-defined rating scale and δ is a positive constant which is used to exclude images having ambiguous aesthetic qualities. The early benchmark CUHK [22] uses a relatively large value of δ to force the two resulting classes distinguishable (only the images associated with the top and bottom 10th percentile of aesthetic scores are retained), while the AVA dataset [37] assigns $\delta = 0$ on the test set, to increase the challenge of binary classification.

The AVA [37] dataset also provides a score distribution for each image based on the available human subjective ratings:

$$\begin{gathered} \mathbf{y}_n = \{y_n^1, y_n^2, ..., y_n^M\}, \\ s.t.\ 0 < y_n^m < 1, \forall m; \sum_{m=1}^{M} y_n^m = 1, \end{gathered} \tag{3}$$

where y_n^m is the percentage of rating scores taking value b_m over all ratings.

B. A Unified Probabilistic Formulation of Aesthetic Labels

As can be seen from (1), (2), and (3), the three different types of aesthetic labels are defined in different forms: one-dimensional binary labels, one-dimensional continuous scores, and multi-dimensional score distributions. Usually, one type of label is used to train an individual IAA model for each task, which is inefficient in practice. In the following we reveal that all of these aesthetic labelling formulas can be unified under a single probabilistic formulation, to consequently leading to a more effective and compact solution to all three IAA tasks.

The binary class label is naturally a probabilistic representation that is associated with a "hard" partition of images. On the standard partition of the AVA dataset ($\delta = 0$), images either "definitely" belong to the high quality set C_{high} or to the low quality set C_{low}. Let $P(\mathbf{x}_n \in C_{high})$ denote the probability of image $\mathbf{x}_n$ belonging to C_{high}, then:

$$\begin{cases} P(\mathbf{x}_n \in C_{high}) = c_n, \\ P(\mathbf{x}_n \in C_{low}) = 1 - c_n. \end{cases} \tag{4}$$

This "hard" partition is highly questionable, since there exists no obvious boundary between image aesthetic levels, and rich information descriptive of aesthetic quality is lost. A more reasonable representation is to use a "soft" set of memberships, which can be achieved via a simple transformation of the average quality score:

$$\begin{cases} P(\mathbf{x}_n \in C_{high}) = \dfrac{\mu_{s_n}}{b_M}, \\ P(\mathbf{x}_n \in C_{low}) = 1 - \dfrac{\mu_{s_n}}{b_M}. \end{cases} \tag{5}$$

In this way, the original average score is scaled to the range $[0, 1]$ and employed to define the "soft" class probabilities. This scale transformation preserves all of the information in μ_{s_n} and brings two benefits. First, the original range of average score μ_{s_n} tends to vary with datasets because of differences in content and the use of different predefined rating scales in $\mathbf{b}$. Scaling them to the same range can simplify the implementation of model training on multiple, diverse datasets. Transforming the original average score into a probabilistic form also makes it possible to train deep IAA models using a more effective loss function, as we will discuss later.

By using (5), quality score regression becomes a probability prediction problem having a two-set partition. The score distribution may naturally be viewed as a more general probabilistic representation having multiple image sets. Specifically, partition the image aesthetic quality levels into M sets $\{C_1, C_2, ..., C_M\}$ containing increasing levels of reported aesthetic quality. By using a "soft" probabilistic representation like (5), each image $\mathbf{x}_n$ will be associated with M discrete probabilities describing its membership to each quality set:

$$P(\mathbf{x}_n \in C_m) = y_n^m, 1 \leq m \leq M. \tag{6}$$

Eqs. (4), (5) and (6) reveal that all three types of aesthetic labels can be derived from a unified probabilistic representation. The only difference lies in the setting of aesthetic quality level and the associated membership. More specifically, the score distribution is the most general form, where multiple (> 2) aesthetic quality levels are employed, and continuous probabilities are used to ascribe the memberships to different quality levels. Reducing the aesthetic quality levels to two (*high* and *low*) leads to the probabilistic formulation of quality score regression. Further quantizing the continuous probabilities to discrete (0 or 1) memberships simplifies the problem to a binary classification task. As a consequence, both the quality score regression and score distribution prediction tasks can be formulated as a probability learning problem. Next we will use this unified probabilistic formulation to motivate us to choose a particularly efficient loss function for training deep IAA models.

C. A More Effective Loss Function

A variety of loss functions have been introduced for IAA model training, including models trained to predict the score distribution [17, 18, 36, 45]. Many of these are too specific to be applied to all three types of aesthetic labels discussed in Section III-B. Moreover, there are "natural pairing" of activation functions and loss functions which result in simpler forms of gradient computation [4]. Given these considerations, we consider the most widely used loss functions which can be adapted to train deep models on three types of aesthetic labels.

The first loss function we consider is the very commonly used MSE loss. The MSE has been used as the baseline loss function on both quality score regression [17, 24] and score distribution prediction problems [18, 36]. It can be easily adapted to optimize learning under our proposed unified probabilistic representation for each of the three types of IAA tasks. The MSE loss has the following general form:

$$\mathcal{L} = \frac{1}{2N} \sum_{n=1}^{N} \sum_{m=1}^{M} \|P(\mathbf{x}_n \in C_m) - t_n^m\|_2^2, \tag{7}$$

where t_n^m is the predicted probability that image $\mathbf{x}_n$ belongs to aesthetic quality set C_m. The MSE loss is usually used in conjunction with linear activation. For the binary case with a two-set partition, the MSE loss simplifies to a single output:

$$\mathcal{L} = \frac{1}{2N} \sum_{n=1}^{N} \|P(\mathbf{x}_n \in C_{high}) - t_n\|_2^2, \tag{8}$$

where t_n is the predicted probability that image $\mathbf{x}_n$ belongs to set C_{high}.

A more robust loss, the Huber loss [16], was introduced in [36] in the context of IAA. It can also be used to learn the probabilistic representation. The general form of the Huber loss is:

$$\mathcal{L} = \begin{cases} \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \frac{1}{2} (P(\mathbf{x}_n \in C_m) - t_n^m)^2, \text{when} |P_n^m - t_n^m| < \epsilon, \\ \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} \epsilon (|P(\mathbf{x}_n \in C_m) - t_n^m| - \frac{1}{2}\epsilon), \text{otherwise}, \end{cases} \tag{9}$$

where P_n^m denotes $P(\mathbf{x}_n \in C_m)$ and ϵ is a constant threshold. The Huber loss is also usually used with linear activation and simplifies to a single output on a two-set partition.

Beyond the MSE and Huber losses which have been widely used to optimize regression tasks, we are more interested in the cross-entropy (CE) loss which is more natural and appropriate when learning probability distributions [35]. The general form of CE loss is:

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} P(\mathbf{x}_n \in C_m) \log t_n^m, \tag{10}$$

where the model output t_n^m follows softmax activation. In the case of a two-set partition, the CE loss degrades to:

$$\begin{aligned} \mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} [&P(\mathbf{x}_n \in C_{high}) \log t_n \\ &+ (1 - P(\mathbf{x}_n \in C_{high})) \log(1 - t_n)], \end{aligned} \tag{11}$$

where t_n is a single output following sigmoid activation. When using a "hard" probability, the above CE loss further simplifies to the logistic loss which is very commonly employed to solve previous binary classification IAA tasks [28, 29, 31, 32].

Using our proposed unified probabilistic formulation, it is natural to use the CE loss to train deep IAA models. Moreover, the CE loss has several advantages over the MSE and Huber losses. Firstly, the CE loss tends to converge faster than the MSE and Huber losses. A qualitative illustration of the three loss functions for the two-set partition problem is shown in Fig. 2. For three exemplar target probabilities $P(\mathbf{x}_n \in C_{high}) = 0.25, 0.5, 0.75$, we plotted the error curves produced by the three loss functions against model output. The sigmoid function always bounds the model output to the range $(0, 1)$, while a linear activation function does not have this property. Thus optimization of a CE loss tends to start at a point closer to the target value than of MSE or Huber

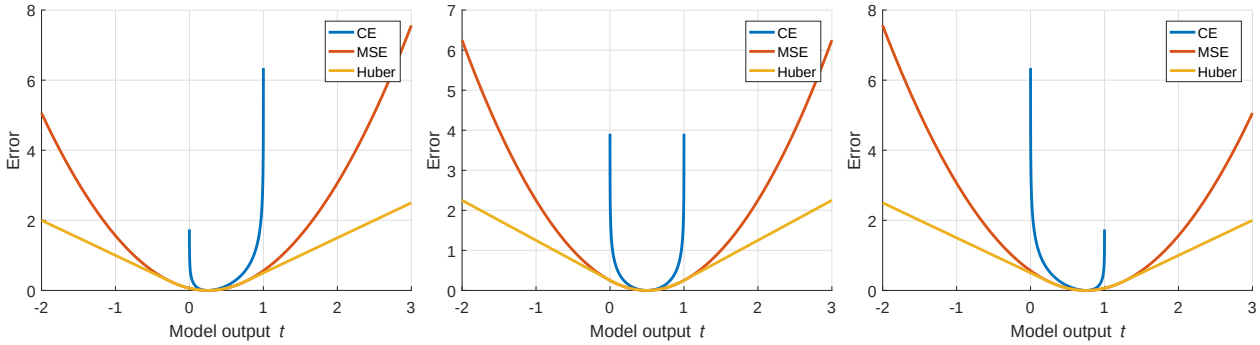


Fig. 2. Illustration of the MSE, Huber and CE losses for target values $0.25, 0.5$ and 0.75 (from left to right). The CE loss is slightly modified into a KL-divergence form to allow the minimum value to be 0, which does not change the optimization process [12].

losses. Note that it is inadvisable to combine the sigmoid (or softmax) activation function with either the MSE or Huber loss since the vanishing gradient problem can occur [11, 49]. More importantly, the CE loss decreases much faster than the MSE and Huber losses when the model output is far from the target value. Finally, as discussed in [4, chapter 6.9, p.235], minimization of the CE loss tends to result in a similar "relative error" for both small and large target values, while the MSE and Huber losses tend to yield a similar "absolute error" for each target. This implies that the CE loss may be expected to perform better when estimating small target values.

On the task of quality score regression, we first train a model using the unified probabilistic formulation, and then we calculate the average quality scores from the predicted probabilities. The regression and classification metrics can then also be evaluated.

D. A Stabilized Score Distribution

Considering the highly subjective nature and inherent ambiguity of image aesthetics, a single ordinary human subject (i.e. without professional training) at any moment is likely to report aesthetic quality scores that are inconsistent with scores s/he may assign the same image at another time, under different environmental, cognitive and emotional conditions. Because of this, the raw score distribution (RSD) associated with a set of human subjective ratings is generally noisy, which reduces the performance of IAA models. Therefore it is greatly desirable to be able to obtain a more reliable score distribution.

Denote by $s_n(k)$ the observed rating score of image $\mathbf{x}_n$ by the kth subject. We can write $s_n(k)$ as follows:

$$s_n(k) = z_n(k) + v_n(k), \tag{12}$$

where $z_n(k)$ is an ideal ground truth subjective aesthetic quality score of $\mathbf{x}_n$ given by the kth subject and $v_n(k)$ models an random noise. Assuming that the noise $v_n(k)$ is independent of $z_n(k)$ (which admitted, is by no means certain) and that it is zero mean with variance $\sigma_{v_n}^2$, then

$$\mu_{s_n} = \mu_{z_n}, \tag{13}$$

and

$$\sigma_{s_n}^2 = \sigma_{z_n}^2 + \sigma_{v_n}^2, \tag{14}$$

where μ_{s_n} and μ_{z_n} are the means of s_n and z_n, and $\sigma_{s_n}^2$ and $\sigma_{z_n}^2$ are the variances of s_n and z_n, respectively.

From (13), μ_{s_n} is an unbiased estimator of the ground-truth average quality score of $\mathbf{x}_n$, but the estimated variance $\sigma_{s_n}^2$ increases with noise. In Fig. 3(a,d,g), we show three two-dimensional visualizations of the RSDs, for images drawn from the AVA dataset having average scores lying in the range $[5, 6]$. As can be seen, the RSDs of different images are interspersed, and there are many outliers. This makes the training of IAA models on the RSD difficult, thereby deteriorating the generalization performance of models trained on them.

Wang *et al.* [45] proposed to fit a Gaussian distribution to the RSD having moments equal to the sample mean μ_{s_n} and sample variance $\sigma_{s_n}^2$, then trained a CNN model using the low-order moments of the fitted Gaussian distribution. However, in the presence of noise, the variance $\sigma_{s_n}^2$ of the observed scores may significantly increase, resulting in broadened distributions, making the fitted Gaussian distribution less discriminative. As shown in Fig. 3(b,e,h), discretizing the fitted Gaussian distribution into the same bins as the RSD makes the spread of probabilities more distinctive but still not sufficiently discriminative.

To ease the learning process and improve the generalization performance of IAA models, we propose a way to form a more stable and distinguishable score distribution with careful consideration of the noise variance $\sigma_{v_n}^2$. Although the $\sigma_{v_n}^2$ is unknown, it is reasonable to assume that $\sigma_{v_n}^2$ is proportional to $\sigma_{s_n}^2$, given that large ambiguities are associated with elevated noise levels. To be more specific, a larger variance of an observed aesthetic score means that the perceptual quality is more controversial. Such images are more likely to confuse the human subjects during annotation, increasing the uncertainty and variance of human decision and score assignment. Thus we have:

$$\sigma_{v_n}^2 = \lambda \sigma_{s_n}^2, \tag{15}$$

where $\lambda \in (0, 1)$. Combining with (14), the $\sigma_{z_n}^2$ can be derived as:

$$\sigma_{z_n}^2 = (1 - \lambda)\sigma_{s_n}^2. \tag{16}$$

We empirically fix $\lambda = 0.5$ throughout this work. To further alleviate the influence of images associated with outlier distributions, i.e., with very large variances, we also place an upper

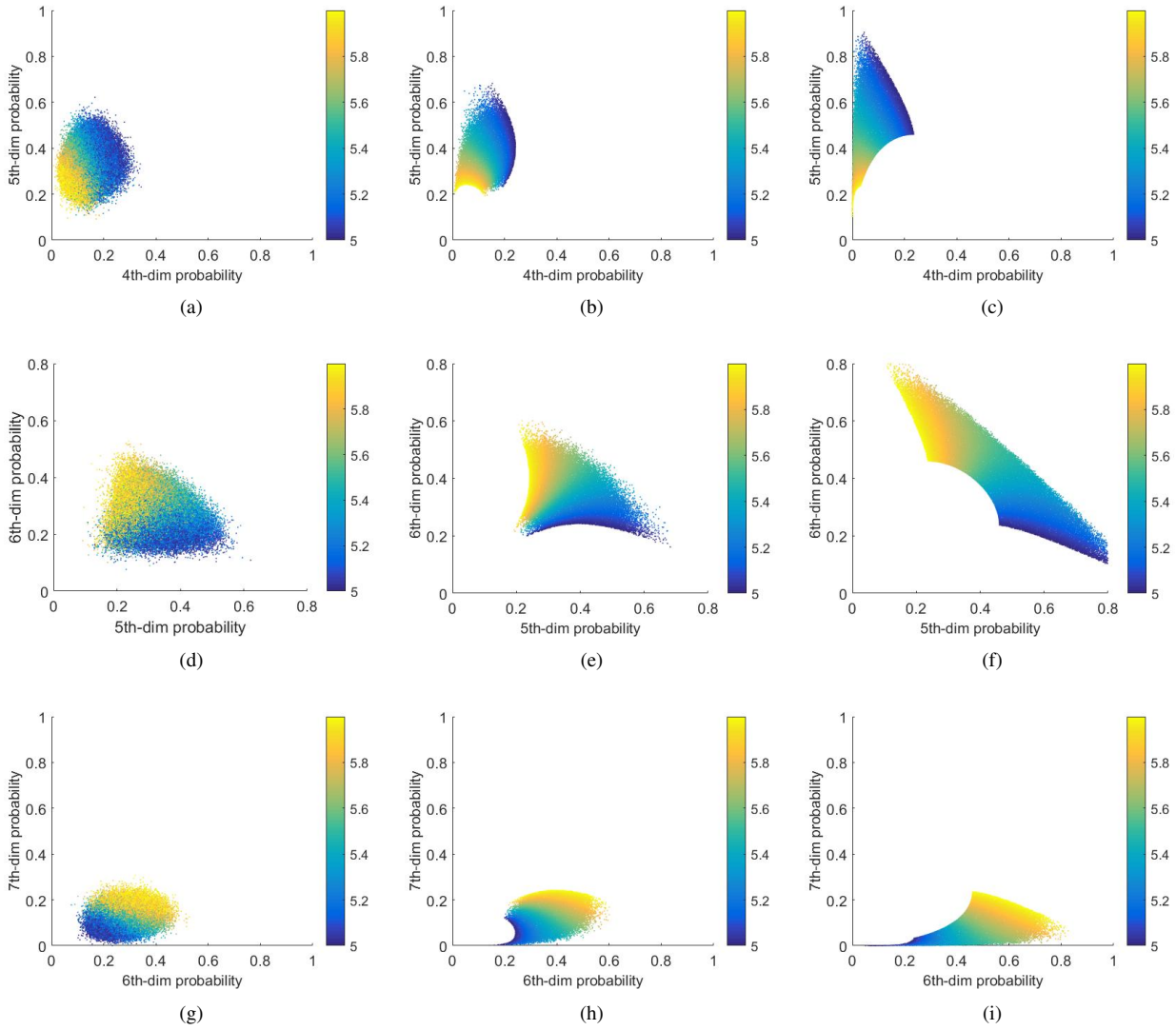


Fig. 3. Visualization of two dimensions (4th and 5th, 5th and 6th, 6th and 7th) of (a,d,g) RSD, (b,e,h) Gaussian distributions with noisy variances, and (c,f,i) our stabilized Gaussian distributions. The images contain all images from the AVA dataset having average quality scores lying in the range $[5, 6]$. The false color represents the values of the average scores.

limit to $\sigma_{z_n}^2$. The distribution of the variances of all the images in AVA after scaling using (16) is depicted in Fig. 4. After the scaling, more than 90% of the images have variances smaller than 1.5, and the remaining 10% of the images have much larger variances. Therefore, we fixed the upper limit to be 1.5 and the ground truth variance $\sigma_{z_n}^2$ is estimated as:

$$\sigma_{z_n}^2 = \min(0.5\sigma_{s_n}^2, 1.5). \tag{17}$$

A Gaussian distribution is thus fitted to the raw score distribution of each image $\mathbf{x}_n$ using μ_{z_n} and $\sigma_{z_n}^2$. The same 2-dimensional visualizations of the stabilized Gaussian distributions are shown in Fig. 3(c,f,i). As can be observed, more distinguishable spreads are obtained on the AVA images. Lastly, we discretize the stabilized Gaussian distribution into the same bins as the RSD and employ the softmax CE loss (10) to train the CNN model. Specifically, we sample M values from the stabilized Gaussian and L_1-normalize the sampled probabilities. The discretized distribution can accurately reproduce the average score with an average reconstruction error smaller than $1e^{-3}$, which is negligible compared to the learning error of the CNN model.

IV. Experiments

A. Datasets

Two representative image aesthetic datasets: AVA [37] and AADB [24] were employed in our experiments.

The **AVA dataset** is the largest publicly available benchmark dedicated to supporting IAA research, as it contains more than 250k labelled images. Each image was rated by, on average, 210 people with integer scores ranging from 1 to 10. For the standard partition of binary classification, images having average scores > 5 are treated as positive examples and images with average scores ≤ 5 are viewed as negative ones. Following common practice on this dataset [29, 32, 37], we employed $\sim$230k images to train the CNN model and the remaining $\sim$20k images to evaluate the performance of

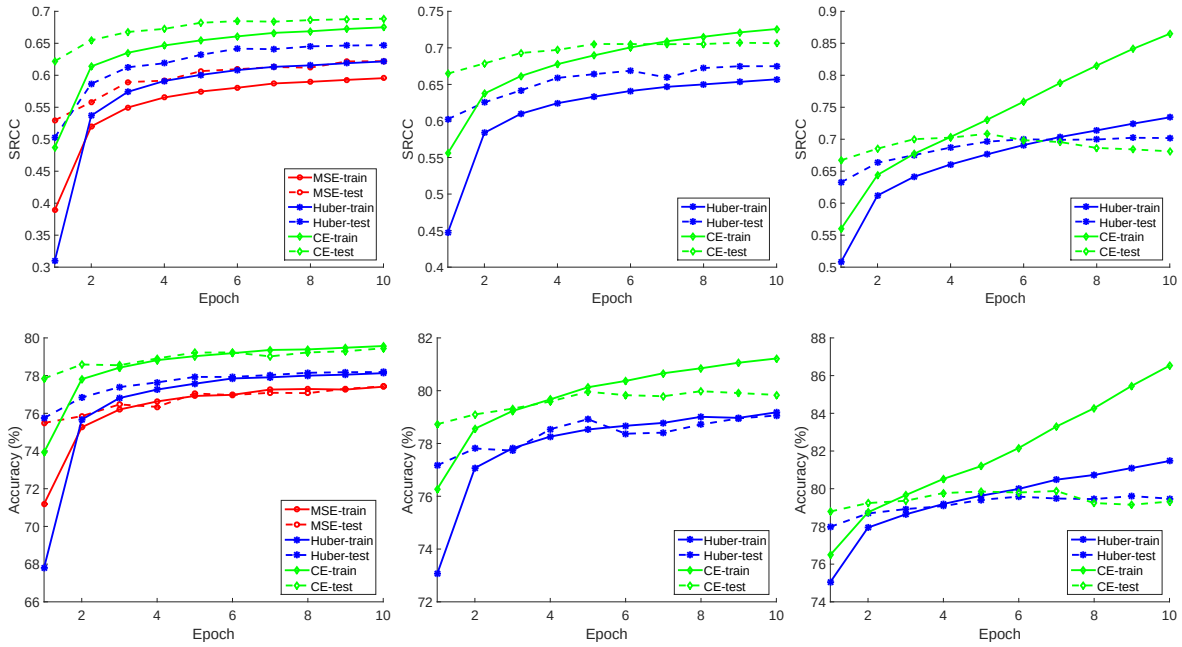


Fig. 5. Plot of the learning curves of the MSE, Huber and CE loss functions against learning rate, which was exponentially decreased between (a) $10^{[-2.5,-3.5]}$, (b) $10^{[-2,-3]}$ and (c) $10^{[-1.5,-2.5]}$. The average quality score was used to fine-tune the ResNet50 model. The MSE loss failed to converge for the last two sets of learning rates. Larger values imply better performance.

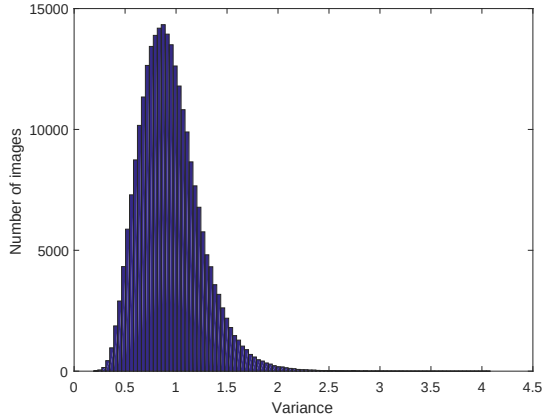


Fig. 4. Distributions of the variances of all images in AVA after the scaling by (16).

competing IAA models. All of the three types of aesthetic labels (binary category, average quality score, and score distribution) are available on this dataset. Six different metrics were used to evaluate model performances on the three IAA tasks: binary classification accuracy for binary label classification; Spearman's rank correlation coefficient (SRCC), Pearson correlation coefficient (PCC) and mean squared error (MSE) for quality score regression/ranking; KL-divergence and the Earth Mover's Distance (EMD) [15] for score distribution prediction.

The **AADB dataset** [24] is a more recent dataset which is specifically designed for aesthetic ranking and attribute learning. It consists of a total of 10,000 images. We followed the standard partition, using 8,500 images for training, 500 images for validation and the remaining 1,000 images for testing. Each image in this database was annotated by, on average, 5 people with integer scores ranging from 1 to 5. We discarded about 100 images from the training set which have only one rating (i.e., the variance cannot be computed). The SRCC was reported on this database following the database creators.

B. Implementation Details

Two representative CNN architectures (the VGG16 [42] and ResNet50 [14]) were employed to implement the proposed models. Both the VGG16 and ResNet50 networks were first pre-trained on the ImageNet [38] classification task, then fine-tuned on the target aesthetic datasets. The last layer was replaced by a new layer adaptive to different aesthetic representations, using the loss functions evaluated in this work. For brevity and fair comparison, all the other operations and settings before the last layer were held fixed over all the different methods.

The input images were resized to a fixed resolution of 384×384, and only random horizontal flip was used for data augmentation. The MatConvNet toolbox [44] was employed to fine-tune the CNN models using stochastic gradient decent with momentum. Throughout our experiments, the batch size and momentum were fixed at 32 and 0.9 for both models. The weight decay was fixed at $1e^{-4}$ and $5e^{-4}$ for ResNet50 and VGG16, respectively. The only parameter that was tuned for each method was the learning rate, in order to account for the different characteristics of the aesthetic representations and loss functions.

TABLE I
BEST PERFORMANCE OBTAINED BY DIFFERENT LOSS FUNCTIONS TRAINED ON RSD. ↑ MEANS LARGER IS BETTER AND ↓ MEANS SMALLER IS BETTER.

Loss	Accuracy (%) ↑	SRCC ↑	PCC ↑	MSE ↓	KL-Div ↓	EMD ↓
MSE	77.87	0.6287	0.6357	0.3368	0.1434	0.0725
EMD	79.33	0.6904	0.6942	0.2932	0.1085	0.0673
Huber	79.32	0.6929	0.6920	0.2915	0.1132	0.0675
CE	**79.78**	**0.7025**	**0.7049**	**0.2842**	**0.1020**	**0.0664**

C. Learning with Quality Scores

As mentioned before, reliable score distributions are much harder to obtain than binary labels or quality scores. Therefore, in practice, it is important to be able to train an IAA model using only binary labels or quality scores. We begin by studying the performance of our model using three loss functions, MSE, Huber and CE, when learning quality scores. To do this, we separately trained a ResNet50 model on the AVA dataset using each of the three loss functions.

The original quality scores were scaled into "soft" probabilities for the two-set partition (refer to (5)). Both classification and regression performances were evaluated for each loss. Following [36], we fixed $\epsilon = \frac{1}{9}$ for the Huber loss. Each loss function was fine-tuned over 10 epochs using an exponentially decreasing learning rate. Three sets of learning rates in the intervals $10^{[-2.5,-3.5]}$, $10^{[-2,-3]}$ and $10^{[-1.5,-2.5]}$ were evaluated for each loss. The learning curves when regressing the average quality scores are plotted in Fig. 5. However, the MSE loss failed to converge for two of the three sets of learning rates, because of the unbounded model output and gradient. Thus, corresponding learning curves for the MSE loss are not included in Fig. 5.

Figure 5 clearly shows that the CE loss achieved the fastest convergence speed among the three competing loss functions, over all three sets of learning rates. Specifically, when using the CE loss, performance on the test set always converged within 5 epochs. When using relatively small learning rates (see Fig. 5(a)), all three loss functions exhibited some degree of under-fitting, where the testing error was similar to or smaller than the training error. By appropriately increasing the learning rate, this problem can be solved, and the model converged to a better local minimum for both the CE and Huber losses, as shown in Fig. 5(b). Further increasing the learning rate would result in over-fitting and worse performance on the test set, as shown in Fig. 5(c).

It is also apparent that the CE loss obtained the best performance among the three loss functions for all three sets of learning rates. Given an appropriate set of learning rates, the Huber loss was also able to deliver competitive performance, yet slightly worse than that achieved using the CE loss. The MSE loss yielded the worst performance and even failed to converge on the two sets of larger learning rates. This may partly explain the moderate SRCC values reported in [24], where the MSE loss was employed to learn a deep model. These results further validate the conclusions reached in the analysis in Section III-C, that the CE loss delivers more efficient and effective performance than its counterparts. It is important to note that it is our unified probabilistic formulation that makes it possible to use the CE loss to train on the quality scores.

TABLE II
COMPARISON OF USING DIFFERENT AESTHETIC LABELS TO TRAIN A DEEP AESTHETIC MODEL ON THE AVA DATASET. ALL OF THE COMPETING METHODS EMPLOYED THE CE LOSS.

Aesthetic labels	SRCC	Accuracy (%)
Binary labels	0.6760	79.45
Quality score	0.7070	79.96
RSD	0.7025	79.78
Gaussian-NV	0.7081	79.95
Gaussian-MV	**0.7141**	**80.35**

D. Learning with Raw Score Distributions

We then compared the performance of models trained using different loss functions when learning raw score distributions using six different evaluation metrics. Along with the MSE loss, Huber loss and CE loss, we also compared performance when using the EMD loss [15, 43], which is specifically designed for distribution learning. As with the experiments on learning quality scores, we tested three sets of learning rates for each loss function. The best results for each loss function are summarized in Table I. Again, the MSE loss obtained the worst performance among all competitors. The EMD and Huber losses achieved comparable results, while the CE loss performed the best on all six metrics. Note that the EMD loss is based on the L_2-distance, which suffers from the same instability and inefficiency as the MSE loss.

E. Learning with Different Aesthetic Labels

This section evaluates IAA performance when using different aesthetic labels. We fixed the loss function to CE, because of its outstanding performance over all of the previous evaluations. Five different aesthetic labels, including binary category labels, average quality scores, RSD, fitted Gaussian with noisy variance (Gaussian-NV) and the proposed fitted Gaussian with modified variance (Gaussian-MV) were compared. We tuned the learning rate for each type of aesthetic label. The SRCC and classification accuracy were employed as the performance metrics. The best results obtained for each type of label are reported in Table II.

It may be seen that the binary category labels achieved the worst performance, especially in terms of the ranking metric SRCC. This is not surprising, since binary labels contain only a small amount of information regarding relative subjective judgments of aesthetic levels. During our experiments, we

TABLE III
COMPARISONS WITH STATE-OF-THE-ART IAA MODELS ON THE AVA DATASET. ↑ MEANS LARGER IS BETTER AND ↓ MEANS SMALLER IS BETTER. THE METHOD MARKED BY § AGGREGATES MULTIPLE LEARNED IMAGE CROPS. THE NIMA RESULTS MARKED BY ‡ WERE TAKEN FROM THE ORIGINAL PAPER, WHICH USED DIFFERENT TRAINING AND TESTING SPLITS THAN THE STANDARD SETTING. THE NIMA RESULTS MARKED BY † WERE OBTAINED BY US USING THE STANDARD PARTITION. THE SYMBOL "–" MEANS RESULTS ARE UNAVAILABLE FOR CORRESPONDING METHODS.

IAA tasks	Classification	Score regression			Distribution prediction		Std. regression		Training cost
Metrics	Accuracy (%)	SRCC ↑	PCC↑	MSE ↓	KL-Div.↓	EMD ↓	SRCC ↑	PCC↑	(Epoch / GPU day)
Murray *et al.* [37]	67.00	–	–	–	–	–	–	–	– / –
Lu *et al.* [28]	74.46	–	–	–	–	–	–	–	– / –
Lu *et al.* [29]	75.41	–	–	–	–	–	–	–	– / –
Kong *et al.* [24]	77.33	0.558	–	–	–	–	–	–	– / –
Mai *et al.* [32] (VGG16)	77.40	–	–	–	–	–	–	–	– / –
Ma *et al.* [31] (VGG16)§	**82.50**	–	–	–	–	–	–	–	>20 / –
Kao *et al.* [21] (ResNet50)	79.08	–	–	–	–	–	–	–	– / –
Deng *et al.* [9] (VGG16)	78.72	–	–	–	–	–	–	–	– / –
Jin *et al.* [18] (GoogLeNet)	80.08	–	–	–	–	–	–	–	∼90 / 3
NIMA [43] (VGG16)‡	80.60	0.592	0.610	–	–	0.052	0.202	0.205	>20 / –
NIMA [43] (Inception-v2)‡	81.51	0.612	0.636	–	–	**0.050**	0.218	0.233	>20 / –
NIMA [43] (ResNet50)†	79.33	0.690	0.694	0.293	0.109	0.067	0.212	0.218	>20 / –
Murray *et al.* [36] (VGG16)	78.70	0.665	–	0.311	0.128	0.067	–	–	– / 12
Murray *et al.* [36] (ResNet101)	80.30	0.709	–	0.279	0.103	0.061	–	–	– / 12
Ours (VGG16)	79.81	0.699	0.701	0.289	0.104	0.067	**0.266**	**0.275**	10 / 1
Ours (ResNet50)	80.35	0.714	0.716	0.276	0.102	0.066	0.247	0.254	**5 / 0.5**
Ours (ResNet101)	80.81	**0.719**	**0.720**	**0.275**	**0.101**	0.065	0.241	0.247	5 / 2

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART IAA MODELS ON THE AADB DATASET.

Methods	SRCC
Kong *et al.* [24]	0.6782
Hou *et al.* [15]	0.6889
Malu *et al.* [33]	0.6890
NIMA [43]	0.7081
Ours (Quality score)	0.7152
Ours (Gaussian-MV)	**0.7264**

TABLE V
CROSS DATASET EVALUATION. SRCC IS REPORTED.

Training set	AVA		AADB	
Testing set	AVA	AADB	AVA	AADB
Kong *et al.* [24]	0.5154	0.3191	0.1566	0.6782
NIMA [43]	0.6904	0.4732	0.2950	0.7081
Ours (Quality score)	0.7070	0.5006	0.3171	0.7152
Ours (Gaussian-MV)	**0.7141**	**0.5176**	**0.3225**	**0.7264**

found that the model trained using binary labels converged faster but also tended to over-fit, even when using a small learning rate. The model trained on average quality score regression not only yielded much better ranking performance but also led to higher classification accuracy. Training on the noisy RSD produced slightly worse performance than regressing on the average quality scores, revealing the unreliability of the RSD. As compared with RSD, the fitted Gaussian distributions performed better with respect to both the classification and regression metrics. Specifically, as compared with the RSD, the proposed Gauss-MV improved the SRCC by more than 0.01. It should be mentioned that this improvement is nontrivial, since our baseline performance was already very competitive, and improving SRCC on ∼20k testing images is very challenging.

F. Comparison with the State-of-the-Art

We now compare the performance of our proposed deep IAA models against several state-of-the-art methods on both the AVA [37] and AADB [24] datasets. The results are listed in Tables III and IV, respectively. The employed CNN models and the corresponding training costs are also reported for comparison. Since many previous studies only conducted binary classification on the AVA dataset, and since the source codes of those works have not been released, we leave their results blank in Table III. It is important to observe that the NIMA method [43] was not created using the standard partition of training and testing images on the AVA dataset, making it impossible to compare their results against the other methods. Therefore, we re-implemented the NIMA method, and tested it using the standard AVA partition. The results of both our implementation and that of the original NIMA paper are included in Table III. It is worth mentioning that both our re-implementation (using ResNet50, obtains EMD 0.067) and the third-party implementation on github [1] (using Inception-ResNet V2, reports EMD 0.07) produced very similar results to each other, but very different results than those reported in the original NIMA paper (using Inception V2, reports EMD 0.05).

Table III shows that the proposed unified solution achieves state-of-the-art performance with respect to most of performance metrics. Our model is particularly effective on the regression/ranking problems and converges much faster than the compared prior ones with significantly improved training efficiency. It is worth mentioning that a variety of specific techniques, such as multi-crop aggregation [29, 31], composition preserving pooling [32], multi-model integration [9, 28, 32] and multi-task learning [21, 24], were used by some of the compared methods to improve their performances on the AVA dataset. Our models outperforms previous ones by

[1] https://github.com/titu1994/neural-image-assessment

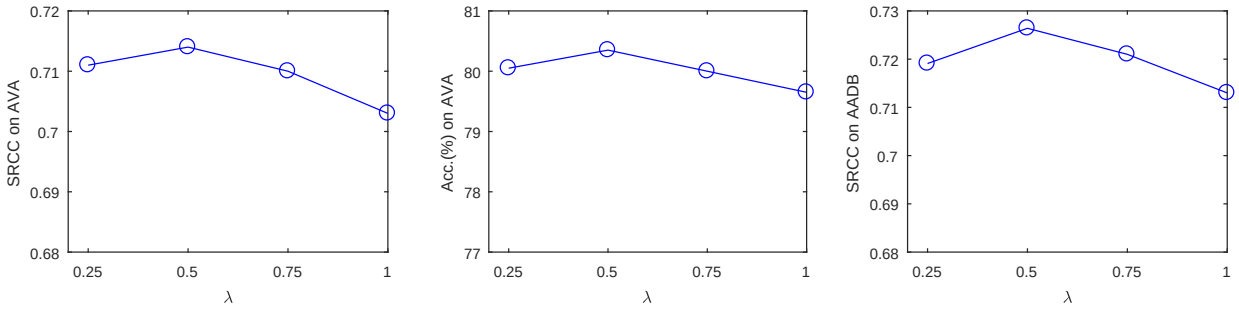


Fig. 6. Effects of different choices of λ on AVA and AADB datasets.

using a very simple framework. The best binary classification accuracy was achieved by the method reported in [31], which deployed a complex saliency detection method and a carefully designed optimization strategy to select multiple image crops. An additional attribute graph was further integrated into their model to further improve performance. In comparison, our proposed probabilistic representation and loss function are free of any such feature extraction processes. We also noticed that VGG16 obtained better results on the task of std. regression than the ResNet models. This is mainly because the standard deviations (2nd order statistics) of images are much harder to learn than their average scores (1st order statistics). Therefore, deeper models such as ResNet101 may over-fit the problem if the training data is not sufficiently large.

The AADB dataset is a new dataset, and only a few results have been reported on it. We implemented the recent method NIMA [43] on this database using the ResNet50 model. Both of the compared methods [24, 33] employed a multi-task learning strategy with additional labels, while [15] aggregated 8 different CNN models. From Table IV, we can see that when using the same ResNet50 model as [33], the NIMA outperforms prior methods, yet our method obtains the best results. Using the Gaussian-MV instead of quality scores again yields a measurable improvement, further validating the generality of our proposed representation.

G. Cross Dataset Evaluation

Given the large variations in reported image aesthetic scores, an effective and practical model should be expected to demonstrate reasonable generalization performance across different sources of data. Towards analyzing the ability of our model in this regard, we followed [24] to conduct a cross dataset evaluation on the AVA and AADB datasets. Since the rating scales on these two datasets are different, we cannot conduct binary classification or score distribution prediction across them. We thus only conducted the quality score regression task, and report the SRCC of the different training and testing settings in Table V. NIMA [43] was implemented by ourselves using the RSD of the database, since the trained model is not available. As compared with [24], both NIMA and our method substantially improve the cross dataset evaluation performance for all settings. Employing the proposed Gaussian-MV further improves performance relative to using the quality scores. An interesting observation is that the model trained on the large scale AVA dataset is more effectively transferable to the small AADB dataset than the reverse. This suggests that training on larger dataset improves the generalization ability of learned IAA models.

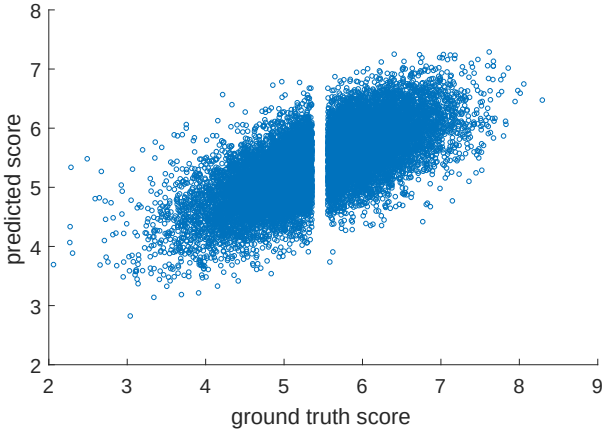


Fig. 7. Scatter plot of ground truth scores versus predicted scores in the ∼20k test images of the AVA dataset. The empty range is caused by the ground truth scores while our model makes continuous predictions.

H. Selection of λ

We conducted an ablation study of the influence of the value of λ in Eq. 15 on model performance. Specifically, we evaluated four even-spaced values λ = 0.25, 0.5, 0.75 and 1.0 on ResNet50 model, with all the other settings fixed. The resulting SRCC and accuracy curves on the AVA dataset and the SRCC on the AADB dataset are plotted in Fig. 6. Our experimental results show that our method is insensitive to variations of λ over a wide range centered around 0.5, while a larger values of λ (close to 1) causes worse performance, which is consistent with our noise assumption. It is worth mentioning that the real noise level of any individual rater is unavailable. When λ = 0, the Gaussian-MV is equal to Gaussian-NV which is not discriminative enough, while when λ = 1, the Gaussian-MV degrades to the single quality score, which causes too much information loss. λ = 0.5 happens to obtain a good balance between preserving distribution information and increasing discriminability.

I. Discussion

In the last stage of our experiments, we make a qualitative analysis of the model trained using our proposed Gaussian-MV representation on the aesthetic quality score prediction

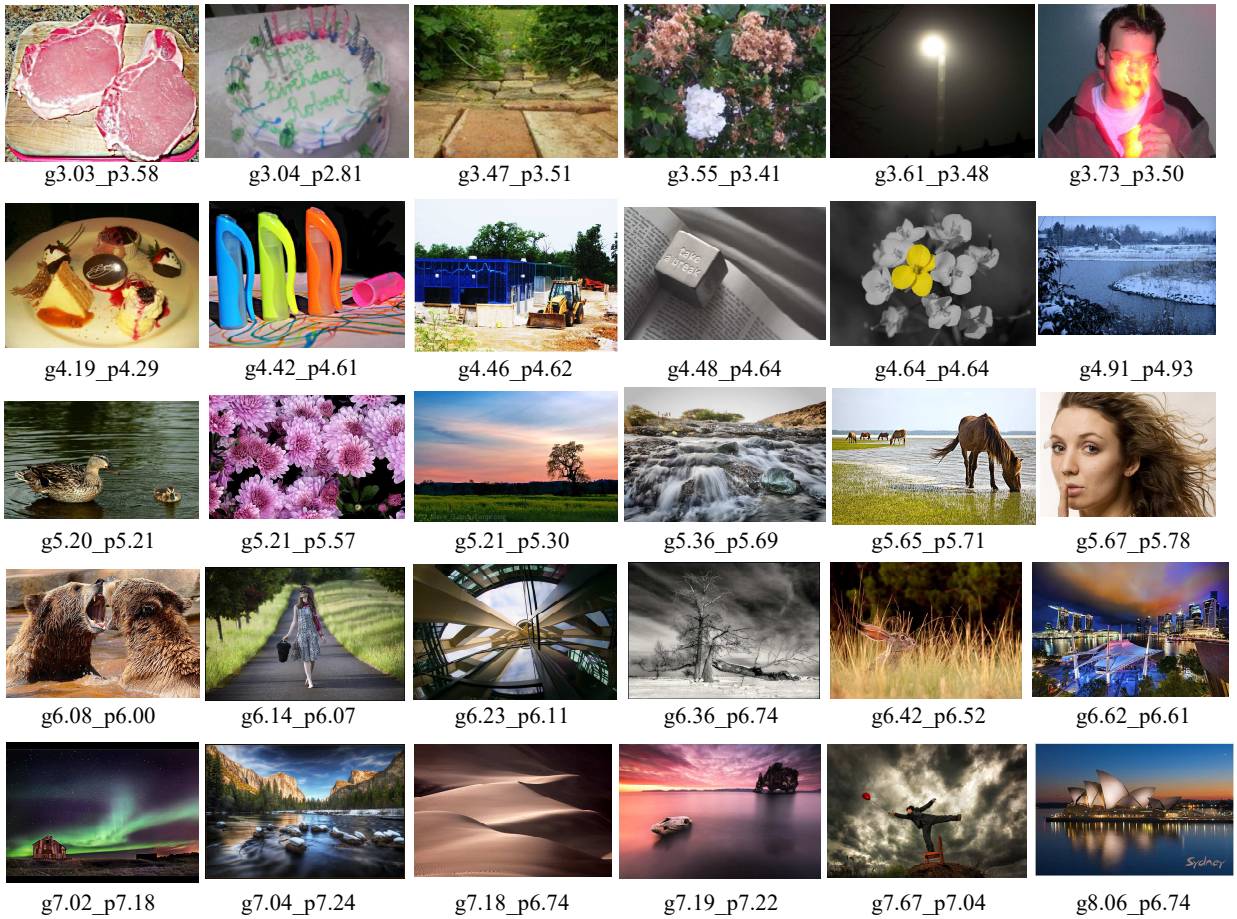


Fig. 8. Examples where our model consistently predicts the ground truth scores. Both the ground truth score (g*.**) and the predicted score (p*.**) are marked for each example.

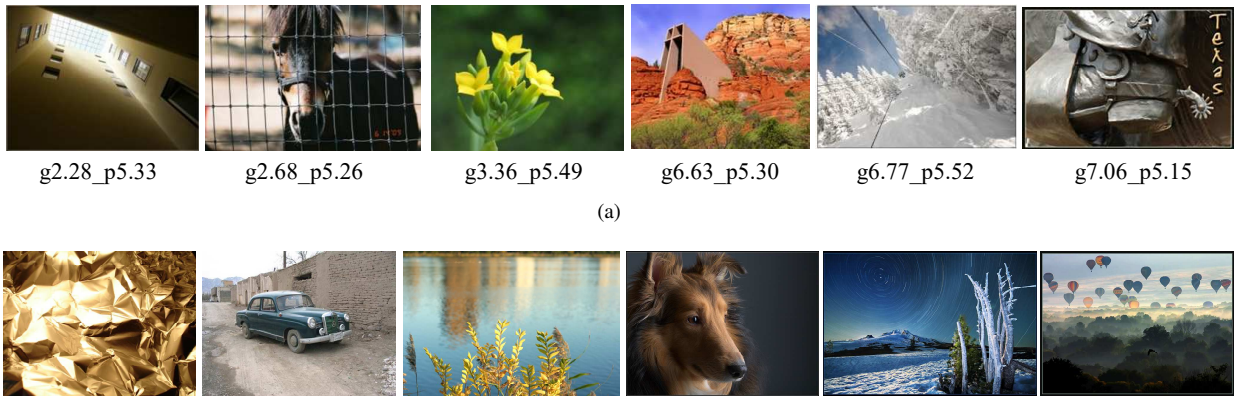


Fig. 9. Examples where our model does not consistently predict the ground truth scores. Both the ground truth score (g*.**) and the predicted score (p*.**) are marked for each example.

problem, because it is more important in practical applications as compared to binary classification and score distribution prediction. A two-dimension scatter plot of the ground truth scores versus the predicted scores on the ∼20k test images of AVA dataset is shown in Fig. 7. As can be seen, the predicted scores of our model are generally linear against the ground truth scores, with a reasonable spread. This indicates that our model is capable to make reliable predictions of image aesthetics on large scale datasets and applications. Some examples that get accurate predictions are shown in Fig. 8. Increases of the predicted aesthetic quality scores are associated with images that are more aesthetically pleasing.

We also found some images whose predicted scores are noticeably inconsistent with the ground truth scores. These

negative examples can be mainly divided into two categories. The first category of examples are generally outliers, where images with moderate aesthetic quality have very low ground truth scores or where ordinary images receive very high ground truth scores. Several examples in this category are shown in Fig. 9(a). The other category contains examples with slightly biased predictions. As shown in Fig. 9(b), the predicted scores for some low quality images may be slightly larger than their associated ground truth scores and conversely for some high quality images. This is mainly caused by an unbalanced distribution of training samples. According to the statistics reported in [37], the ground truth scores of all of the images in the AVA dataset follow a Gaussian distribution, and the scores of most of the images lie in the range of [4,7]. Thus, our model tends to make predictions biased towards the dominant score range. Fortunately, we rarely found that images having low aesthetic quality were predicted by our model to have high aesthetic scores, or vice versa.

V. Conclusion

We have studied the close relationship between the three main tasks that define the image aesthetic assessment (IAA) problems, and we provided a unified probabilistic formulation to handle these different tasks. Using the unified probabilistic representation, the cross-entropy loss was naturally employed to learn deep IAA models. By using our probabilistic formulation, both the efficiency of the training process and the performance of the trained models were much improved. A more reliable aesthetic score distribution was also proposed to train deep IAA models, which further improved the performance on both the binary classification and quality score regression problems. Comprehensive analysis and extensive comparisons were conducted, where we showed that our proposed method achieved state-of-the-art IAA performance on the two available aesthetic benchmarks.

Acknowledgments

This work is supported by China NSFC grant (no. 61672446) and Hong Kong RGC RIF grant (R5001-18).

References

[1] S. A. Amirshahi and J. Denzler. Judging aesthetic quality in paintings based on artistic inspired color features. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017.

[2] T. Ang. *Digital Photographer's Handbook*. Dorling Kindersley Publishing, Incorporated, 2002.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *ACM International Conference on Multimedia*, pages 271–280, 2010.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[6] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301, 2006.

[8] Y. Deng, C. C. Loy, and X. Tang. Aesthetic-driven image enhancement by adversarial learning. *arXiv preprint arXiv:1707.05251*, 2017.

[9] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.

[10] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li. The role of attractiveness in web image search. In *ACM International Conference on Multimedia*, pages 63–72, 2011.

[11] P. Golik, P. Doetsch, and H. Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, volume 13, pages 1756–1760, 2013.

[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] L. Hou, C.-P. Yu, and D. Samaras. Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016.

[16] P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[17] B. Jin, M. V. O. Segovia, and S. Süsstrunk. Image aesthetic predictors based on weighted CNNs. In *IEEE International Conference on Image Processing (ICIP)*, pages 2291–2295, 2016.

[18] X. Jin, L. Wu, C. Song, X. Li, G. Zhao, S. Chen, J. Chi, S. Peng, and S. Ge. Predicting aesthetic score distribution through cumulative Jensen-Shannon divergence. *arXiv preprint arXiv:1708.07089*, 2017.

[19] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, 2011.

[20] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[21] Y. Kao, R. He, and K. Huang. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing*, 26(3):1482–1495, 2017.

[22] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 419–426, 2006.

[23] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik. Deep convolutional neural models for picture-quality prediction. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017.

[24] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679. Springer, 2016.

[25] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, pages 469–478. Wiley Online Library, 2010.

[26] Z. Liu, Z. Wang, Y. Yao, L. Zhang, and L. Shao. Deep active learning with contaminated tags for image aesthetics assessment. *IEEE Transactions on Image Processing*, 2018.

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[28] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia*, 17(11):2021–2034, 2015.

[29] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *IEEE International Conference on Computer Vision*, pages 990–998, 2015.

[30] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399. Springer, 2008.

[31] S. Ma, J. Liu, and C. W. Chen. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] L. Mai, H. Jin, and F. Liu. Composition-preserving deep photo aesthetics assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016.

[33] G. Malu, R. S. Bapi, and B. Indurkhya. Learning photography aesthetics with deep CNNs. *arXiv preprint arXiv:1707.03981*, 2017.

[34] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *IEEE International Conference on Computer Vision*, pages 1784–1791, 2011.

[35] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[36] N. Murray and A. Gordo. A deep architecture for unified aesthetic prediction. *arXiv preprint arXiv:1708.04890*, 2017.

[37] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[39] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *International Conference on World Wide Web*, pages 439–448. ACM, 2012.

[40] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[41] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 879–886. ACM, 2018.

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[43] H. Talebi and P. Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[44] A. Vedaldi and K. Lenc. MatConvNet – convolutional neural networks for MATLAB. In *Proceeding of the ACM International Conference on Multimedia*, 2015.

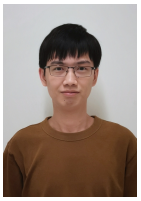
[45] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang. Image aesthetics assessment using deep Chatterjee's machine. In *International Joint Conference on Neural Networks (IJCNN)*, pages 941–948, 2017.

[46] O. Wu, W. Hu, and J. Gao. Learning to predict the perceived visual quality of photos. In *IEEE International Conference on Computer Vision*, pages 225–232, 2011.

[47] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)*, 35(2):11, 2016.

[48] V. Yanulevskaya, J. Uijlings, E. Bruni, A. Sartori, E. Zamboni, F. Bacci, D. Melcher, and N. Sebe. In the eye of the beholder: employing statistical analysis and eye tracking for analyzing abstract paintings. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 349–358. ACM, 2012.

[49] P. Zhou and J. Austin. Learning criteria for training neural network classifiers. *Neural Computing & Applications*, 7(4):334–342, 1998.



Hui Zeng is currently a Ph.D candidate in the Department of Computing, The Hong Kong Polytechnic University. He received his B.Sc. and M.Sc. degrees from the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, in 2014 and 2016, respectively. His research interests include image processing and computer vision.

Zisheng Cao is currently with imaging group of DJI. He received his B.S. and M.S. degrees from Tsinghua University, in 2005 and 2007, respectively, and PhD degree from the University of Hong Kong, in 2014. His research interests include image signal processing and machine learning. Before joining in DJI, he was a research scientist in Philips.

Lei Zhang (F'18) received his B.Sc. degree in 1995 from Shenyang Institute of Aeronautical Engineering, Shenyang, P.R. China, and M.Sc. and Ph.D degrees in Control Theory and Engineering from Northwestern Polytechnical University, Xian, P.R. China, in 1998 and 2001, respectively. From 2001 to 2002, he was a research associate in the Department of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Department of Computing, The Hong Kong Polytechnic University, as an Assistant Professor. Since July 2017, he has been a Chair Professor in the same department. His research interests include Computer Vision, Image and Video Analysis, Pattern Recognition, and Biometrics, etc. Prof. Zhang has published more than 200 papers in those areas. As of 2019, his publications have been cited more than 43,000 times in literature. Prof. Zhang is a Senior Associate Editor of IEEE Trans. on Image Processing, and is/was an Associate Editor of SIAM Journal of Imaging Sciences, IEEE Trans. on CSVT, and Image and Vision Computing, etc. He is a Clarivate Analytics Highly Cited Researcher from 2015 to 2018. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cslzhang/.

Al Bovik (F'95) is the Cockrell Family Regents Endowed Chair Professor at The University of Texas at Austin. His research interests include image processing, digital television, digital streaming video, and visual perception. For his work in these areas he has been the recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from the Optical Society of America, a 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has also received about 10 best journal paper awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. A Fellow of the IEEE, his recent books include The Essential Guides to Image and Video Processing. He co-founded and was longest-serving Editor-in-Chief of the IEEE Transactions on Image Processing, and also created/Chaired the IEEE International Conference on Image Processing which was first held in Austin, Texas, 1994.